PATENT APPLICATION

METHODS OF GENETIC ANALYSIS

Inventors: Michael Mittman, a citizen of the United States,
             Residing at:
             2377 St. Francis Drive
             Palo Alto, California 94303

             David Mack, a citizen of the United States,
             Residing at:
             2076 Monterey Avenue
             Menlo Park, California 94025

             David Lockhart, a citizen of the United States
             Residing at:
             510 Torrey Point Rd.
             Del Mar, California 92014

Assignee: Affymetrix, Inc.

METHODS OF GENETIC ANALYSIS

CROSS REFERENCE TO RELATED APPLICATIONS

5   The present application is a non-provisional application claiming priority from

Provisional U.S. Patent Application Serial No. 60/100,678 filed September 17, 1998.

BACKGROUND OF THE INVENTION

The present invention provides a unique pool of nucleic acid sequences useful for

analyzing molecular interactions of biological interest. The invention therefore relates to

10   diverse fields impacted by the nature of molecular interaction, including chemistry, biology,

medicine, and medical diagnostics.

FIELD OF THE INVENTION

Many biological functions are carried out by regulating the expression levels of various

genes, either through changes in levels of transcription (*e.g.* through control of initiation,

15   provision of RNA precursors, RNA processing, *etc.*) of particular genes, through changes in the

copy number of the genetic DNA, or through changes in protein synthesis.   For example,

control of the cell cycle and cell differentiation, as well as diseases, are characterized by the

variations in the transcription levels of a group of genes.

Gene expression is not only responsible for physiological functions, but also associated

20   with pathogenesis. For example, the lack of sufficient functional tumor suppressor genes and/or

the over expression of oncogene/protooncogenes leads to tumorgenesis. (See, e.g., Marshall,

Cell, 64: 313-326 (1991) and Weinberg, Science, 254: 1138-1146 (1991.))   Thus, changes in the

expression levels of particular genes (*e.g.* oncogenes or tumor suppressors), serve as signposts for the presence and progression of various diseases.

As a consequence, novel techniques and apparatus are needed to study gene expression in specific biological systems.

5   All documents, i.e., publications and patent applications, cited in this disclosure, including the foregoing, are incorporated by reference herein in their entireties for all purposes to the same extent as if each of the individual documents were specifically and individually indicated to be so incorporated by reference herein in its entirety.

SUMMARY OF THE INVENTION

10   The invention provides nucleic acid sequences which are complementary to particular genes and makes them available for a variety of analyses, including, for example, gene expression analysis. For example, in one embodiment the invention comprises an array comprising of any 10 or more, 100 or more, 1000, or more, 10,000 or more or 100,000 or more nucleic acid probes containing 9 or more consecutive nucleotides from the sequences listed in

15   SEQ ID NOS: 1 -127811, or the perfect match, perfect mismatch, antisense match or antisense mismatch thereof. In a further embodiment, the invention comprises the use of any of the above arrays or fragments disclosed in TABLE 1 to: monitor gene expression levels by hybridization of the array to a DNA library; monitor gene expression levels by hybridization to an mRNA-protein fusion compound; identify polymorphisms; identify biallelic markers; produce genetic

20   maps; analyze genetic variation; comparatively analyze gene expression between different species; analyze gene knockouts; or, to hybridize tag-labeled compounds. In a further embodiment the invention comprises a method of analysis comprising of hybridizing one or

more pools of nucleic acids to two or more of the fragments disclosed in TABLE 1 and

detecting said hybridization. In a further embodiment the invention comprises the use of any

one or more of the fragments disclosed in TABLE 1 as a primer for PCR. In a further

embodiment the invention comprises the use of any one or more of the fragments disclosed in

5   TABLE 1 as a ligand.

DETAILED DESCRIPTION OF THE INVENTION

I. Definitions

10   Massive Parallel Screening: The phrase "massively parallel screening" refers to the

simultaneous screening of at least about 100, preferably about 1000, more preferably about

10,000 and more preferably about 100,000 different nucleic acid hybridizations.

          Nucleic Acid: The terms "nucleic acid" or "nucleic acid molecule" refer to a

deoxyribonucleotide or ribonucleotide polymer in either single-or double-stranded form, and

15   unless otherwise limited, would encompass analogs of natural nucleotides that can function in a

similar manner as naturally occurring nucleotides. Nucleic acids may be derived from a variety

or sources including, but not limited to, naturally occurring nucleic acids, clones, synthesis in

solution or solid phase synthesis.

          Probe: As used herein a "probe" is defined as a nucleic acid capable of binding to a target

20   nucleic acid of complementary sequence through one or more types of chemical bonds, usually

through complementary base pairing, usually through hydrogen bond formation. As used

herein, a probe may include natural (*i.e.* A, G, U, C, or T) or modified bases (7-deazaguanosine,

inosine, *etc.*). In addition, the bases in probes may be joined by a linkage other than a

phosphodiester bond, so long as it does not interfere with hybridization. Thus, probes may be

peptide nucleic acids in which the constituent bases are joined by peptide bonds rather than

phosphodiester linkages.

5   Target nucleic acid: The term "target nucleic acid" or "target sequence" refers to a

nucleic acid or nucleic acid sequence which is to be analyzed. A target can be a nucleic acid to

which a probe will hybridize. The probe may or may not be specifically designed to hybridize

to the target. It is either the presence or absence of the target nucleic acid that is to be detected,

or the amount of the target nucleic acid that is to be quantified. The term target nucleic acid

10   may refer to the specific subsequence of a larger nucleic acid to which the probe is directed or to

the overall sequence (*e.g.*, gene or mRNA) whose expression level it is desired to detect. The

difference in usage will be apparent from context.

mRNA or transcript: The term "mRNA" refers to transcripts of a gene. Transcripts are

RNA including, for example, mature messenger RNA ready for translation, products of various

15   stages of transcript processing. Transcript processing may include splicing, editing and

degradation.

Subsequence: "Subsequence" refers to a sequence of nucleic acids that comprise a part

of a longer sequence of nucleic acids.

Perfect match: The term "match," "perfect match," "perfect match probe" or "perfect

20   match control" refers to a nucleic acid that has a sequence that is perfectly complementary to a

particular target sequence. The nucleic acid is typically perfectly complementary to a portion

(subsequence) of the target sequence. A perfect match (PM) probe can be a "test probe", a

"normalization control" probe, an expression level control probe and the like. A perfect match control or perfect match is, however, distinguished from a "mismatch" or "mismatch probe."

Mismatch: The term "mismatch," "mismatch control" or "mismatch probe" refers to a nucleic acid whose sequence is deliberately selected not to be perfectly complementary to a particular target sequence. As a non-limiting example, for each mismatch (MM) control in a high-density probe array there typically exists a corresponding perfect match (PM) probe that is perfectly complementary to the same particular target sequence. The mismatch may comprise one or more bases. While the mismatch(es) may be located anywhere in the mismatch probe, terminal mismatches are less desirable because a terminal mismatch is less likely to prevent hybridization of the target sequence. In a particularly preferred embodiment, the mismatch is located at or near the center of the probe such that the mismatch is most likely to destabilize the duplex with the target sequence under the test hybridization conditions. A homo-mismatch substitutes an adenine (A) for a thymine (T) and vice versa and a guanine (G) for a cytosine (C) and vice versa. For example, if the target sequence was: AGGTCCA, a probe designed with a single homo-mismatch at the central, or fourth position, would result in the following sequence: TCCTGGT.

Array: An "array" is a solid support with at least a first surface having a plurality of different nucleic acid sequences attached to the first surface.

Gene Knockout: the term "gene knockout, " as defined in Lodish et al *Molecular Cell Biology 3rd Edition*, Scientific American Books pub., which is hereby incorporated in its entirety for all purposes is, is a technique for selectively inactivating a gene by replacing it with a mutant allele in an otherwise normal organism.

DNA Library - as used herein the term "genomic library" or "genomic DNA library" refers to a collection of cloned DNA molecules consisting of fragments of the entire genome (genomic library) or of DNA copies of all the mRNA produced by a cell type (cDNA library) inserted into a suitable cloning vector.

5   Polymorphism - "polymorphism" refers to the occurrence of two or more genetically determined alternative sequences or alleles in a population. A polymorphic marker or site is the locus at which divergence occurs. Preferred markers have at least two alleles, each occurring at a frequency of greater than 1%, and more preferably greater than 10% or 20% of the selected population. A polymorphic locus may be as small as one base pair. Polymorphic markers

10   include restriction fragment length polymorphisms, variable number or tandem repeats (VNTR's), hypervariable regions, minisatellites, dinucleotide repeats, trinucleotide repeats, tetranucleotide repeats, simple sequence repeats, and insertion elements such as ALU. The first identified allelic form is arbitrarily designated as the reference form and other allelic forms are designated as alternative or variant alleles. The allelic form occurring most frequently in a

15   selected population is sometimes referred to as the wildtype form. Diploid organisms may be homozygous or heterozygous for allelic forms. A diallelic or biallelic polymorphism has two forms. A triallelic polymorphism has three forms.

Genetic map - a "genetic map" is a map which presents the order of specific sequences on a chromosome.

20   Genetic variation - "genetic variation" refers to variation in the sequence of the same region between two or more organisms.

Hybridization - the association of two complementary nucleic acid strands or their

derivatives (such as PNA) to form double stranded molecules. Hybrids can contain two DNA

strands, two RNA strands, or one DNA and one RNA strand.

mRNA-protein fusion - a compound whereby an mRNA is directly attached to the

5   peptide or protein it incodes by a stable covalent linkage.

Ligand - any molecule, other than an enzyme substrate, that binds tightly and specifically

to a macromolecule, for example, a protein, forming a macromolecule-ligand complex.

II. General

10   TABLE 1, encompassed in Appendix I, presents target sequences included in the

invention. Each target sequence from columns 2,5 and 8 corresponds to and represents at least

four additional nucleic acid sequences included in the invention. For example, if the first

nucleic acid sequence listed in TABLE 1 is: **cgccggaaagcgtgccaatcaacat** the additional

sequences included in the invention which are represented by this nucleic acid sequence are, for

15   example, :

gcggcctttcgcacggccagttgta = (perfect) sense match

gcggcctttcgctcggccagttgta = sense mismatch

tacaactggccgagcgaaaggccgc = (perfect) antisense match

tacaactggccgtgcgaaaggccgc = antisense mismatch

20   Accordingly, for each nucleic acid sequence listed in TABLE 1, this disclosure includes

the corresponding sense match, sense mismatch, antisense match and antisense mismatch. The

position of the mismatch is not limited to the above example, it may be located anywhere in the nucleic acid sequence and may comprise one or more bases.

Consequently, the present invention includes: a) the target sequences listed in TABLE 1, columns 2, 5 and 8 or the sense-match, sense mismatch, antisense match or antisense

5   mismatch thereof; b) clones which comprise the target nucleic acid sequences listed in TABLE 1, columns 2, 5 and 8 or the sense-match, sense mismatch, antisense match or antisense mismatch thereof; c) longer nucleotide sequences which include the nucleic acid sequences listed in TABLE 1, columns 2, 5, and 8 or the sense-match, sense mismatch, antisense match or antisense mismatch thereof and d) subsequences greater than 9 nucleotides in length of the target

10   nucleic acid sequences listed in TABLE 1, columns 2, 5, and 8 or the sense match, sense mismatch, antisense match or antisense mismatch.

Target sequences were chosen from clusters of known murine genes available on the Unigene database as of August 15, 1996. Target sequences were selected using the computer based methods described in US patent application No. 08/772,376 incorporated herein by

15   reference for all purposes.

For each target sequence listed in TABLE 1 is a corresponding genbank database accession number. These accession numbers allow for the identification of sequences located in the genbank sequence database through the use of computer programs such as BLAST. Access to BLAST is available to the public through the internet at, for example,

20   http://www.ncbi.nim.nih.gov. One of skill in the art will be familiar with the use of the BLAST program to obtain information about particular sequences in order to, for example, determine the species from which the sequence is derived, determine the gene from which the sequence is

derived, to determine other genes and species which contain similar sequences and to determine the degree of similarity between one sequence and another. All information relating to the target sequences available through the genbank database is hereby incorporated by reference for all purposes.

5   The present invention provides a pool of unique nucleotide sequences complementary to murine genes and ESTs in particular embodiments which alone, or in combinations of 2 or more, 10 or more, 100 or more, 1,000 or more, 10,000 or more, or 100,000 or more, can be used for a variety of applications.

In one embodiment, the present invention provides for a pool of unique nucleotide

10   sequences which are complementary to approximately 6500 murine genes formed into a high density array of probes suitable for array based massive parallel gene expression. Array based methods for monitoring gene expression are disclosed and discussed in detail in U.S. Patent No. Applications, 08/670,118, and 08/772,376 and PCT Application WO 92/10588 (published on June 25, 1992), all of which are incorporated herein by reference for all purposes. Generally

15   those methods of monitoring gene expression involve (1) providing a pool of target nucleic acids comprising RNA transcript(s) of one or more target gene(s), or nucleic acids derived from the RNA transcript(s); (2) hybridizing the nucleic acid sample to a high density array of probes and (3) detecting the hybridized nucleic acids and calculating a relative expression (transcription, RNA processing or degradation) level.

20   The development of Very Large Scale Immobilized Polymer Synthesis or VLSIPS™ technology has provided methods for making very large arrays of nucleic acid probes in very small arrays. See U.S. Patent No. 5,143,854 and PCT patent publication Nos. WO 90/15070

and 92/10092, and Fodor *et al., Science,* 251, 767-77 (1991), each of which is incorporated herein by reference. U.S. Patent application Serial No. 08/670,118, describes methods for making arrays of nucleic acid probes that can be used to detect the presence of a nucleic acid containing a specific nucleotide sequence. Methods of forming high density arrays of nucleic

5   acids, peptides and other polymer sequences with a minimal number of synthetic steps are known. The nucleic acid array can be synthesized on a solid substrate by a variety of methods, including, but not limited to, light-directed chemical coupling, and mechanically directed coupling.

In a preferred detection method, the array of immobilized nucleic acids, or probes, is

10   contacted with a sample containing target nucleic acids, to which a flourescent label is attached. Target nucleic acids hybridize to the probes on the array and any non-hybridized nucleic acids are removed. The array containing the hybridized target nucleic acids are exposed to light which excites the flourescent label. The resulting flourescent intensity, or brightness, is detected. Relative brightness is used to determine which probe is the best candidate for the

15   perfect match to the hybridized target nucleic acid because flourescent intensity (brightness) corresponds to binding affinity. Once the position of the perfect match probe is known, the sequence of the hybridized target nucleic is known because the sequence and position of the probe is known.

In the array of the present invention the probes are presented in pairs, one probe in each

20   pair being a perfect match to the target sequence and the other probe being identical to the perfect match probe except that the central base is a homo-mismatch. Mismatch probes provide a control for non-specific binding or cross-hybridization to a nucleic acid in the sample other

than the target to which the probe is directed. Thus, mismatch probes indicate whether a hybridization is or is not specific. For example, if the target is present, the perfect match probes should be consistently brighter than the mismatch probes because fluorescence intensity, or brightness, corresponds to binding affinity. (See, for example US Patent No. 5,324,633, which is incorporated herein for all purposes.) In addition, if all central mismatches are present, the mismatch probes can be used to detect a mutation. Finally the difference in intensity between the perfect match and the mismatch probe (I(PM)-I(MM)) provides a good measure of the concentration of the hybridized material. See pending PCT application No. 98/11223, which is incorporated herein by reference for all purposes. The probe pairs are presented in both sense and antisense orientation, thereby eliciting a total of four probes per target sequence: sense match, sense mismatch, antisense match and antisense mismatch.

In another embodiment, the current invention provides a pool of sequences which may be used as probes for their complementary genes listed in the genbank database. Methods for making probes are well known. See for example Sambrook, Fritsche and Maniatis. "Molecular Cloning A laboratory Manual" 2nd Ed. Cold Spring Harbor Press (1989) ("Maniatis et al.") which is hereby incorporated in its entirety by reference for all purposes. Maniatis et al. describes a number of uses for nucleic acid probes of defined sequence. Some of the uses described by Maniatis et al. include: to screen cDNA or genomic DNA libraries, or subclones derived from them, for additional clones containing segments of DNA that have been isolated and previously sequenced; in Southern, northern, or dot-blot hybridization to identify or detect the sequences of specific genes; in Southern, or dot-blot hybridization of genomic DNA to detect specific mutations in genes of known sequence; to detect specific mutations generated by

site-directed mutagenesis of cloned genes; and to map the 5' termini of mRNA molecules by

primer extensions. Maniatis et al. describes other uses for probes throughout. See also Alberts

et al. *Molecular Biology of the Cell 3rd edition,* Garland Publishing Inc. (1994) p. 307 and

Lodish et al. *Molecular Cell Biology, 3rd edition,* Scientific American Books (1995) p. 285-286,

5   each of which is hereby incorporated by reference in its entirety for all purposes, for a brief

discussion of the use of nucleic acid probes in *in situ hybridization.* Other uses for probes

derived from the sequences disclosed in this invention will be readily apparent to those of skill

in the art. See, for example, Lodish et al. *Molecular Cell Biology, 3rd edition,* Scientific

American Books (1995) p.229-233, incorporated above, for a description of the construction of

10   genomic libraries.

In another embodiment, the current invention may be combined with known methods to

monitor expression levels of genes in a wide variety of contexts. For example, where the effects

of a drug on gene expression are to be determined, the drug will be administered to an organism,

a tissue sample, or a cell and the gene expression levels will be analyzed. For example, nucleic

15   acids are isolated from the treated tissue sample, cell, or a biological sample from the organism

and from an untreated organism tissue sample or cell, hybridized to a high density probe array

containing probes directed to the gene of interest and the expression levels of that gene are

determined. The types of drugs that may be used in these types of experiments include, but are

not limited to, antibiotics, antivirals, narcotics, anti-cancer drugs, tumor suppressing drugs, and

20   any chemical composition which may affect the expression of genes *in vivo* or *in vitro.* The

current invention is particularly suited to be used in the types of analyses described by, for

example, pending US Applications No. 08/772,376 and PCT Application No. 98/11223, each of

which is incorporated by reference in its entirety for all purposes. As described in Wodicka et

al., Nature Biotechnology 15 (1997), (hereby incorporated by reference in its entirety for all

purposes), because mRNA hybridization correlates to gene expression level, hybridization

patterns can be compared to determine differential gene expression. As non-limiting examples:

5   hybridization patterns from samples treated with certain types of drugs may be compared to

hybridization patterns from samples which have not been treated or which have been treated

with a different drug; hybridization patterns for samples infected with a specific virus may be

compared against hybridization patterns from non-infected samples; hybridization patterns for

samples with cancer may be compared against hybridization patterns for samples without

10   cancer; hybridization patterns of samples from cancerous cells which have been treated with a

tumor suppressing drug may be compared against untreated cancerous cells, etc. Zhang et al.,

Science 276 1268-1272, (hereby incorporated by reference in its entirety for all purposes),

provides an example of how gene expression data can provide a great deal of insight into cancer

research. One skilled in the art will appreciate that a wide range of applications will be

15   available using 2 or more, 10 or more, 100 or more, 1000 or more, 10,000 or more or 100,000 or

more of the Table 1 sequences as probes for gene expression analysis. The combination of the

DNA array technology and the mouse specific probes in this disclosure is a powerful tool for

studying gene expression.

In another embodiment, the invention may be used in conjunction with the techniques

20   which link specific proteins to the mRNA which encodes the protein. (See for example Roberts

and Szostak Proc. Natl, Acad. Sci. 94 12297-12302 (1997) which is incorporated herein in its

entirety for all purposes.) Hybridization of these mRNA-protein fusion compounds to arrays

comprised of 2 or more, 10 or more, 100 or more, 1000 or more, 10,000 or more, or 100,000 or more the sequences disclosed in the present invention provides a powerful tool for monitoring expression levels.

In one embodiment, the current invention provides a pool of unique nucleic acid sequences which can be used for parallel analysis of gene expression under selective conditions. Without wishing to be limited, genetic selection under selective conditions could include: variation in the temperature of the organism's environment; variation in pH levels in the organism's environment; variation in an organism's food (type, texture, amount etc.); variation in an organism's surroundings; etc. Arrays, such as those in the present invention, can be used to determine whether gene expression is altered when an organism is exposed to selective conditions.

Methods for using nucleic acid arrays to analyze genetic selections under selective conditions are known. (See for example, R. Cho et al., Proc. Natl. Acad. Sci. *95* 3752-3757 (1998) incorporated herein in its entirety for all purposes.) Cho et al. describes the use of a high-density array containing oligonucleotides complementary to every gene in the yeast *Saccharomyces cerevisiae* to perform two-hybrid protein-protein interaction screens for *S. cerevisiae* genes implicated in mRNA splicing and microtubule assembly. Cho et al. was able to characterize the results of a screen in a single experiment by hybridization of labeled DNA derived from positive clones. Briefly, as described by Cho et al., two proteins are expressed in yeast as fusions to either the DNA-binding domain or the activation domain of a transcription factor. Physical interaction of the two proteins reconstitutes transcriptional activity, turning on a gene essential for survival under selective conditions. In screening for novel protein-protein

interactions, yeast cells are first transformed with a plasmid encoding a specific DNA-binding

fusion protein. A plasmid library of activation domain fusions derived from genomic DNA is

then introduced into these cells. Transcriptional activation fusions found in cells that survive

selective conditions are considered to encode peptide domains that may interact with the DNA-

5   binding domain fusion protein. Clones are then isolated from the two-hybrid screen and mixed

into a single pool. Plasmid DNA is purified from the pooled clones and the gene inserts are

amplified using PCR. The DNA products are then hybridized to yeast whole genome arrays for

characterization. The methods employed by Cho et al. are applicable to the analysis of a range

of genetic selections. High density arrays created using two or more, 10 or more, 100 or more,

10   1000 or more, 10,000 or more, or 100,000 or more of the sequences disclosed in the current

invention can be used to analyze genetic selections in the mouse system using the methods

described in Cho et al.

In another embodiment, the current invention provides a pool of unique nucleic acid

sequences which can be used to identify biallelic markers, providing a novel and efficient

15   approach to the study of genetic variation. For example, methods for using high density arrays

comprised of probes which are complementary to the genomic DNA of a particular species to

interrogate polymorphisms are well known. (See for example, US Patent Application Nos.

08/965,620 and 08/853,370 which are hereby incorporated herein for all purposes.) Pools of 2

or more, 10 or more, 100 or more, 1000 or more, 10,000 or more, or 100,000 or more of the

20   sequences disclosed in this invention combined with the methods described in the above patent

applications provides a tool for studying genetic variation in the mouse system.

In another embodiment of the invention, genetic variation can be used to produce genetic maps of various strains of mouse. Winzeler et al., "Direct Allelic Variation Scanning of the Yeast Genome" Science (in press) (1998), which is hereby incorporated for all purposes describe methods for conducting this type of screening with arrays containing probes

5   complementary to the yeast genome. Briefly, genomic DNA from strains which are phenotypically different are isolated, fragmented, and labelled. Each strain is then hybridized to identical arrays comprised of the nucleic acid sequences complementary to the system being studied. Comparison of hybridization patterns between the various strains then serve as genetic markers. As described by Winzler et al, these markers can then be used for linkage analysis.

10   High density arrays created from 2 or more, 10 or more, 100 or more, 1000 or more, 10,000 or more, or 100,000 or more of the sequences disclosed in this invention can be used to study genetic variation using the methods described by Winzler et al.

In another embodiment, the present invention may be used for cross-species comparisons. One skilled in the art will appreciate that it is often useful to determine whether a

15   gene present in one species, for example the mouse, is present in a conserved format in another species, including, without limitation, mouse, human, chicken, zebrafish, *drosophila*, or yeast. See, for example, Andersson et al., Mamm Genome 7(10):717-734 (1996,) which is hereby incorporated by reference for all purposes, which describes the utility of cross-species comparisons. The use of 2 or more, 10 or more, 100 or more, 1000 or more, 10,000 or more or

20   100,000 or more of the sequences disclosed in this invention in an array can be used to determine whether any of the sequence from one or more of the murine genes represented by the sequences disclosed in this invention is conserved in another species by, for example,

hybridizing genomic nucleic acid samples from another species to an array comprised of the sequences disclosed in this invention. Areas of hybridization will yield genomic regions where the nucleotide sequence is highly conserved between the interrogation species and the mouse.

In another embodiment, the present invention may be used to characterize the genotype

5   of knockouts. Methods for using gene knockouts to identify a gene are well known. See for example, Lodish et al. *Molecular Cell Biology, 3rd Edition*, Scientific American Books pub pp. 292-296 and US Patent No. 5,679,523 which are hereby incorporated by reference for all purposes. By isolating genomic nucleic acid samples from knockout species with a known phenotype and hybridizing the samples to an array comprised of 2 or more, 10 or more, 100 or

10   more, 1000 or more, 10,000 or more, or 100,000 or more of the sequences disclosed in this invention, candidates genes which contribute to the phenotype will be identified and made accessible for further characterization.

In another embodiment, the present invention may be used to identify new gene family members. Methods of screening libraries with probes are well known. (See, for example,

15   Maniatis et al, incorporated by reference above.) Because the present invention is comprised of nucleic acid sequences from specific known genes, 2 or more, 10 or more, 100 or more, 1000 or more, 10,000 or more, or 100,000 or more of sequences disclosed in this invention may be used as probes to screen genomic libraries to look for additional family members of those genes from which the target sequences are derived.

20   In another embodiment, the present invention may be used to provide nucleic acid sequences to be used as tag sequences. Tag sequences are a type of genetic "bar code" which can be used to label compounds of interest. The analysis of deletion mutants using tag

sequences is described in, for example, Shoemaker et al., Nature Genetics 14 450-456 (1996), which is hereby incorporated by reference in its entirety for all purposes. Shoemaker et al. describes the use of PCR to generate large numbers of deletion strains. Each deletion strain is labelled with a unique 20-base tag sequence that can be hybridized to a high-density

5   oligonucleotide array. The tags serve as unique identifiers (molecular bar codes) that allow analysis of large numbers of deletion strains simultanously through selective growth conditions. The use of tag sequences need not be limited to this example however. The utility of using unique known short oligonucleotide sequences capable of hybridizing to a nucleic acid array to label various compounds will be apparent to one skilled in the art. One or more, 10 or more,

10   100 or more, 1000 or more, 10,000 or more, or 100,000 or more of the Table 1 sequences are excellent candidates to be used as tag sequences.

In another embodiment of the invention, the sequences of this invention may be used to generate primers directed to their corresponding genes as disclosed in the Genbank or any other public database. These primers may be used in such basic techniques as sequencing or PCR, see

15   for example Maniatis et al., incorporated by reference above.

In another embodiment, the invention provides a pool of nucleic acid sequences to be used as ligands for specific genes. The sequences disclosed in this invention may be used as ligands to their corresponding genes as disclosed in the Genbank or any other public database. Compounds which specifically bind known genes are of interest for a variety of uses. One

20   particular clinical use is to act as an antisense protein which specifically binds and disables a gene which has been, for example, linked to a disease. Methods and uses for ligands to specific

genes are known. See for example, US Patent No. 5,723,594 which is hereby incorporated by reference in its entirety for all purposes.

In a preferred embodiment, the hybridized nucleic acids are detected by detecting one or more labels attached to the sample nucleic acids. The labels may be incorporated by any of a number of means well known to those of skill in the art. In one embodiment, the label is simultaneously incorporated during the amplification step in the preparation of the sample nucleic acids. Thus, for example, polymerase chain reaction (PCR) with labeled primers or labeled nucleotides will provide a labeled amplification product. In a another embodiment, transcription amplification, as described above, using a labeled nucleotide (*e.g.* fluorescein-labeled UTP and/or CTP) incorporates a label into the transcribed nucleic acids.

Alternatively, a label may be added directly to the original nucleic acid sample (*e.g.*, mRNA, polyA mRNA, cDNA, *etc.*) or to the amplification product after the amplification is completed. Means of attaching labels to nucleic acids are well known to those of skill in the art and include, for example nick translation or end-labeling (*e.g.* with a labeled RNA) by kinasing of the nucleic acid and subsequent attachment (ligation) of a nucleic acid linker joining the sample nucleic acid to a label (*e.g.*, a fluorophore).

Detectable labels suitable for use in the present invention include any composition detectable by spectroscopic, photochemical, biochemical, immunochemical, electrical, optical or chemical means. Useful labels in the present invention include biotin for staining with labeled streptavidin conjugate, magnetic beads (*e.g.*, Dynabeads™), fluorescent dyes (*e.g.*, fluorescein, texas red, rhodamine, green fluorescent protein, and the like), radiolabels (*e.g.*, ^{3}H, ^{125}I, ^{35}S, ^{14}C, or ^{32}P), phosphorescent labels, enzymes (*e.g.*, horse radish peroxidase, alkaline phosphatase and

others commonly used in an ELISA), and colorimetric labels such as colloidal gold or colored glass or plastic (*e.g.*, polystyrene, polypropylene, latex, etc.) beads. Patents teaching the use of such labels include U.S. Patent Nos. 3,817,837; 3,850,752; 3,939,350; 3,996,345; 4,277,437; 4,275,149; and 4,366,241, each of which is hereby incorporated by reference in its entirety for

5 all purposes.

Means of detecting such labels are well known to those of skill in the art. Thus, for example, radiolabels may be detected using photographic film or scintillation counters, fluorescent markers may be detected using a photodetector to detect emitted light. Enzymatic labels are typically detected by providing the enzyme with a substrate and detecting the reaction

10 product produced by the action of the enzyme on the substrate, and colorimetric labels are detected by simply visualizing the colored label.

The label may be added to the target nucleic acid(s) prior to, or after the hybridization. So called "direct labels" are detectable labels that are directly attached to or incorporated into the target nucleic acid prior to hybridization. In contrast, so called "indirect labels" are joined to

15 the hybrid duplex after hybridization. Often, the indirect label is attached to a binding moiety that has been attached to the target nucleic acid prior to the hybridization. Thus, for example, the target nucleic acid may be biotinylated before the hybridization. After hybridization, an aviden-conjugated fluorophore will bind the biotin bearing hybrid duplexes providing a label that is easily detected. For a detailed review of methods of labeling nucleic acids and detecting

20 labeled hybridized nucleic acids see *Laboratory Techniques in Biochemistry and Molecular Biology, Vol. 24: Hybridization With Nucleic Acid Probes*, P. Tijssen, ed. Elsevier, N.Y., (1993) which is hereby incorporated by reference in its entirety for all purposes.

Fluorescent labels are preferred and easily added during an *in vitro* transcription reaction. In a preferred embodiment, fluorescein labeled UTP and CTP are incorporated into the RNA produced in an *in vitro* transcription reaction as described above.

EXAMPLE

The following example serves to illustrate the type of experiment that could be conducted using the invention.

Expression Monitoring by Hybridization to High Density Oligonucleotide Arrays

Arrays containing the desired number of probes can be synthesized using the method described in US Patent No. 5,143,854, incorporated by reference above. Extracted poly (A)$^+$RNA can then be coverted to cDNA using the methods described below. The cDNA is then transcribed in the presence of labeled ribonucleotide triphosphates. The label may be biotin or a dye such as fluorescein. RNA is then fragmented with heat in the presence of magnesium ions. Hybridizations are carried out in a flow cell that contains the two-dimensional DNA probe arrays. Following a brief washing step to remove unhybridized RNA, the arrays are scanned using a scanning confocal microscope.

1. A method of RNA preparation:

Labeled RNA is prepared from clones containing a T7 RNA polymerase promoter site by incorporating labeled ribonucleotides in an IVT reaction. Either biotin-labeled or fluorescein-labeled UTP and CTP (1:3 labeled to unlabeled) plus unlabeled ATP and GTP is used for the reaction with 2500 U of T7 RNA polymerase. Following the reaction unincorporated nucleotide triphosphates are removed using size-selective membrane such as Microcon - 100, (Amicon, Beverly, MA). The total molar concentration of RNA is based on a measurement of the absorbance at 260 nm. Following quantitation of RNA amounts, RNA is fragmented randomly to an average length of approximately 50 bases by heating at 94° in 40 mM Tris-acetate pH 8.1, 100 mM potassium acetate, 30 mM magnesium acetate, for 30 to 40 min. Fragmentation reduces possible interference from RNA secondary structure, and minimizes the effects of multiple interactions with closely spaced probe molecules. For material

made directly from cellular RNA, cytoplasmic RNA is extracted from cells by the method of

Favaloro et al. Methods Enzymol. 65:718-749 (1980) hereby incorporated by reference for all

purposes, and poly (A)$^+$ RNA is isolated with an oligo dT selection step using, for example,

Poly Atract, (Promega, Madison, WI). RNA can be amplified using a modification of the

5   procedure described by Eberwine et al. Proc. Natl. Acad. Sci. USA 89:3010-3014 (1992) hereby

incorporated by reference for all purposes. Microgram amounts of poly (A)$^+$ RNA are converted

into double stranded cDNA using a cDNA synthesis kit (kits may be obtained from Life

Technologies, Gaithersburg, MD) with an oligo dT primer incorporating a T7 RNA polymerase

promoter site. After second-strand synthesis, the reaction mixture is extracted with

10   phenol/chloroform, and the double-stranded DNA isolated using a membrane filtration step

using, for example, Microcon -100, (Amicon). Labeled cRNA can be made directly from the

cDNA pool with an IVT step as described above. The total molar concentration of labeled

cRNA is determined from the absorbance at 260nm and assuming an average RNA size of 1000

ribonucleotides. The commonly used convention is that 1 OD is equivalent to 40 ug of RNA,

15   and that 1 ug of cellular mRNA consists of 3 pmol of RNA molecules. Cellular mRNA may

also be labeled directly without any intermediate cDNA synthesis steps. In this case, Poly (A)$^+$

RNA is fragmented as described, and the 5' ends of the fragments are kinased and then

incubated overnight with a biotinylated oligoribonucleotide (5'-biotin-AAAAAA-3') in the

presence of T4 RNA ligase (available from Epicentre Technologies, Madison, WI).

20   Alternatively, mRNA has been labeled directly by UV-induced cross-linking to a psoralen

derivative linked to biotin (available from Schleicher & Schuell, Keene, NH).

2. Array hybridization and Scanning:

        Array hybridization solutions can be made containing 0.9 M NaCl, 60mM EDTA, and

0.005% Triton X-100, adjusted to pH 7.6 (referred to as 6xSSPE-T). In addition, the solutions

25   should contain 0.5 mg/ml unlabeled, degraded herring sperm DNA (available from Sigma, St.

Louis, MO). Prior to hybridization, RNA samples are heated in the hybridization solution to

99°C for 10 min, placed on ice for 5 min, and allowed to equilibrate at room temperature before

being placed in the hybridization flow cell. Following hybridization, the solutions are removed,

the arrays washed with 6xSSPE-T at 22C for 7 min, and then washed with 0.5xSSPE-T at 40°C

30   for 15 min. When biotin labeled RNA is used the hybridized RNA should be stained with a

streptavidin-phycoerythrin in 6xSSPE-T at 40°C for 5 min. The arrays are read using a
scanning confocal microscope made by Molecular Dynamics (commercially available through
Affymetrix, Santa Clara, CA). The scanner uses an argon ion laser as the excitation source, with
the emission detected by a photomultiplier tube through either a 530 nm bandpass filter

5   (flourescein) or a 560 nm longpass filter (phycoerythrin). Nucleic acids of either sense or
antisense orientations may be used in hybridization experiments. Arrays for probes with either
orientation (reverse complements of each other) are made using the same set of photolithgraphic
masks by reversing the order of the photochemical steps and incorporating the complementary
nucleotide.

10   3. Quantitative analysis of hybridization patterns and intensities.

Following a quantitative scan of an array, a grid is aligned to the image using the known
dimensions of the array and the corner control regions as markers. The image is then reduced to
a simple text file containing position and intensity information using software developed at
Affymetrix (available with the confocal scanner). This information is merged with another text

15   file that contains information relating physical position on the array to probe sequence and the
identity of the RNA (and the specific part of the RNA) for which the oligonucleotide probe is
designed. The quantitative analysis of the hybridization results involves a simple form of
pattern recognition based on the assumption that, in the presence of a specific RNA, the perfect
match (PM) probes will hybridize more strongly on average than their mismatch (MM) partners.

20   The number of instances in which the PM hybridization is larger than the MM signal is
computed along with the average of the logarithm of the PM/MM ratios for each probe set.
These values are used to make a decision (using a predefined decision matrix) concerning the
presence or absence of an RNA. To determine the quantitative RNA abundance, the average of
the difference (PM-MM) for each probe family is calculated. The advantage of the difference

25   method is that signals from random cross-hybridization contribute equally, on average, to the
PM and MM probes, while specific hybridization contributes more to the PM probes. By
averaging the pairwise differences, the real signals add constructively while the contributions
from cross-hybridization tend to cancel. When assessing the differences between two different
RNA samples, the hybridization signals from side-by-side experiments on identically

30   synthesized arrays are compared directly. The magnitude of the changes in the average of the

difference (PM-MM) values is interpreted by comparison with the results of spiking experiments as well as the signals observed for the internal standard bacterial and phase RNAs spiked into each sample at a known amount. Data analysis programs, such as those described in US Patent No.08/828,952 perform these operations automatically.

5

CONCLUSION

The inventions herein provide a pool of unique nucleic acid sequences which are complementary to approximately 6500 specific known murine genes. These sequences can be used for a variety of types of analyses.

10   The above description is illustrative and not restrictive. Many variations of the invention will become apparent to those of skill in the art upon review of this disclosure. The scope of the invention should, therefore, be determined not with reference to the above description, but instead be determined with reference to the appended claims along with their full scope of equivalents.

15